Understanding Bias in Nonlinear Panel Models: Some Recent Developments

Manuel Arellano and Jinyong Hahn

Econometric Society World Congress
University College London, 23 August 2005

Introduction

- Review recently developed bias-adjusted estimation methods for nonlinear panel data models with fixed effects.
- For some models, like static linear and logit regressions, there exist fixed-T consistent estimators as $n \to \infty$.
- Fixed T consistency is a desirable property because for many panels T is much smaller than n.
- However, these type of estimators are not available in general, and when they are, their properties do not normally extend to estimates of average marginal effects.
- Moreover, the common parameters of certain nonlinear fixed effects models are unidentified in a fixed T setting, so that fixed-T consistent point estimation is not possible.
- In other cases, fixed-T consistent estimation at the standard root-n rate is impossible.

- The number of periods available for many household, firm-level or country panels is such that it is not less natural to talk of time-series finite sample bias than of fixed-T inconsistency or underidentification.
- In this light, an alternative reaction to the fact that micro panels are short is to ask for approximately unbiased estimators as opposed to estimators with no bias at all.
- That is, estimators with biases of order $1/T^2$ as opposed to the standard magnitude of $1/T$.
- This alternative approach has the potential of overcoming some of the fixed-T identification difficulties and the advantage of generality.

Outline

1. Fixed effects estimation and the incidental parameters problem
2. Bias-correction of the estimator.
   (a) Formulae for the order $1/T$ bias
   (b) Estimators of the bias
   (c) Infinitely iterated bias-correction
3. Bias-correction of the moment equation.
4. Bias-correction of the concentrated likelihood.
5. Other approaches leading to bias correction: Cox and Reid's and Lancaster's approaches based on orthogonality, and their extensions.
6. Quasi maximum likelihood estimation for dynamic models.
7. Estimation of marginal effects.
8. Automatic methods based on simulation.

1. Incidental Parameters Problem with Large T

- Let the data be $z_{it} = (y_{it}, x_{it})$, $(t = 1, ..., T; i = 1, ..., n)$, where y_{it} is the dependent variable, and x_{it} is a strictly exogenous variable. Let θ be a common parameter, α_i an individual effect, and
$$\prod_{t=1}^{T} f\left(y_{it} \mid x_{it}, \theta_0, \alpha_{i0}\right)$$
a density of $y_{i1}, \ldots, y_{iT}$ conditional on $x_{i1}, \ldots, x_{iT}$, assuming that y_{it} are independent across i and t.

- The fixed effects estimator is obtained by doing ML treating each α_i as a parameter to be estimated. Concentrating out the α_i leads to:

$$\widehat{\theta} \equiv \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f\left(y_{it} \mid \theta, \widehat{\alpha}_i(\theta)\right), \quad \widehat{\alpha}_i(\theta) \equiv \underset{\alpha}{\operatorname{argmax}} \sum_{t=1}^{T} \log f\left(y_{it} \mid \theta, \alpha\right).$$

- Let
$$L(\theta) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\left[\sum_{t=1}^{T} \log f\left(y_{it} \mid \theta, \widehat{\alpha}_i(\theta)\right)\right].$$

- It follows from the usual extremum estimator properties that as $n \to \infty$ with T fixed, $\widehat{\theta} = \theta_T + o_p(1)$, where $\theta_T \equiv \operatorname{argmax}_\theta L(\theta)$. In general,
$$\theta_T \neq \theta_0.$$

- This is the incidental parameters problem noted by Neyman and Scott (1948). The source of this problem is the estimation error of $\widehat{\alpha}_i(\theta)$.

Example 1

- Consider a simple model where
$$y_{it} \overset{i.i.d.}{\sim} \mathcal{N}\left(\alpha_{i0}, \sigma_0^2\right), (t = 1, ..., T; i = 1, ..., n),$$
or
$$\log f\left(y_{it}; \sigma^2, \alpha_i\right) = C - \frac{1}{2}\log\sigma^2 - \frac{\left(y_{it} - \alpha_i\right)^2}{2\sigma^2}.$$

- Here, we may write $\theta = \sigma^2$, and the MLE is such that
$$\begin{aligned} \widehat{\alpha}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} \equiv \overline{y}_i, \\ \widehat{\theta} &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i\right)^2. \end{aligned}$$

- It is straightforward to show that
$$\widehat{\theta} = \theta_0 - \frac{1}{T}\theta_0 + o_p(1)$$
as $n \to \infty$ with T fixed.

- In this example, the bias is easy to fix by equating the denominator with the correct degrees of freedom $n(T-1)$.

- The bias should be small for large enough T, i.e.,
$$\lim_{T\to\infty} \theta_T = \theta_0.$$
- Furthermore, for smooth likelihoods we usually have
$$\theta_T = \theta_0 + \frac{B}{T} + O\left(\frac{1}{T^2}\right)$$
for some B.
- In Example 1, $B = -\theta_0$.
- We will also generally have, as $n, T \to \infty$,
$$\sqrt{nT}\left(\widehat{\theta} - \theta_T\right) \xrightarrow{d} N\left(0, \Omega\right)$$
for some Ω.
- Under these general conditions the fixed effects estimator is asymptotically biased even if T grows at the same rate as n.
- For $n/T \to \rho$, say,
$$\sqrt{nT}\left(\widehat{\theta} - \theta_0\right) = \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\frac{B}{T} + O\left(\sqrt{\frac{n}{T^3}}\right) \xrightarrow{d} N\left(B\sqrt{\rho}, \Omega\right).$$
- Thus, even when T grows as fast as n, asymptotic confidence intervals based on the fixed effects estimator will be incorrect, due to the limiting distribution of $\sqrt{nT}\left(\widehat{\theta} - \theta_0\right)$ not being centered.

- Similar to the bias of the fixed effects estimand $\theta_T - \theta_0$, we can also expand in orders of magnitude of T:
  - the bias in the expected fixed effects score at θ_0
$$E\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\log f\left(y_{it} \mid \theta_0, \widehat{\alpha}_i\left(\theta_0\right)\right)\right] = \frac{1}{T}b_i\left(\theta_0\right) + o\left(\frac{1}{T}\right)$$
  - and the bias in the expected concentrated likelihood at an arbitrary θ
$$E\left[\frac{1}{T}\sum_{t=1}^{T}\log f\left(y_{it} \mid \theta, \widehat{\alpha}_i\left(\theta\right)\right) - \frac{1}{T}\sum_{t=1}^{T}\log f\left(y_{it} \mid \theta, \overline{\alpha}_i\left(\theta\right)\right)\right] = \frac{1}{T}\beta_i\left(\theta\right) + o\left(\frac{1}{T}\right)$$
    where $\overline{\alpha}_i\left(\theta\right)$ maximizes
$$\lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=1}^{T}\log f\left(y_{it} \mid \theta, \alpha\right)\right].$$
- These expansions motivate alternative approaches to bias correction based on
  - adjusting the estimator,
  - the estimating equation, or
  - the objective function.
- We next discuss these three approaches in turn.

2. Bias-Correction of the Estimator

- An analytical bias correction is to plug into the formula for B estimators of its unknown components to construct $\widehat{B}$, and then form a bias corrected estimator

$$\widehat{\theta}^{1} \equiv \widehat{\theta} - \frac{\widehat{B}}{T}.$$

Formulae for the Order $1/T$ Bias

- In order to implement this idea, we need to have an explicit formula for B. For this purpose, it is convenient to define

$$\begin{aligned} u_{it}(\theta,\alpha) &\equiv \frac{\partial}{\partial\theta}\log f(y_{it}|\theta,\alpha), \quad v_{it}(\theta,\alpha) \equiv \frac{\partial}{\partial\alpha_i}\log f(y_{it}|\theta,\alpha), \\ V_{2it}(\theta,\alpha) &= v_{it}^2(\theta,\alpha) + v_{it}^{\alpha_i}(\theta,\alpha), \\ U_{it}(\theta,\alpha) &\equiv u_{it}(\theta,\alpha) - v_{it}(\theta,\alpha)\, E\left[v_{it}^{\alpha_i}\right]^{-1} E\left[u_{it}^{\alpha_i}\right], \quad \mathcal{I}_i \equiv -E\left[\frac{\partial U_{it}(\theta_0,\alpha_{i0})}{\partial\theta'}\right]. \end{aligned}$$

  Note that $E\left[U_{it}^{\alpha_i}\right] = 0$, which means that U_{it} and v_{it} are orthogonalized.

- We denote the derivative with respect to θ or α_i by appropriate superscripts, and for convenience we suppress the arguments when expressions are evaluated at true values.

- It can be shown that

$$B = \left(\lim_{n\to\infty}\frac{1}{n}\sum\nolimits_{i=1}^{n}\mathcal{I}_i\right)^{-1}\lim_{n\to\infty}\frac{1}{n}\sum\nolimits_{i=1}^{n} b_i(\theta_0)$$

  where $b_i(\theta_0)/T$ is the $1/T$ bias of the score function.

- $b_i(\theta_0)$ is given by:

$$b_i(\theta_0) = -\frac{E\left[v_{it}U_{it}^{\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]} + \frac{E\left[U_{it}^{\alpha_i\alpha_i}\right]E\left[v_{it}^2\right]}{2\left(E\left[v_{it}^{\alpha_i}\right]\right)^2}.$$

- To obtain this expression, we expand $T^{-1}\sum_{t=1}^T u_{it}(\theta_0, \widehat{\alpha}_i(\theta_0))$ around α_{i0}, combine it with a stochastic expansion of the fixed effects estimation error at the truth

$$\widehat{\alpha}_i(\theta_0) - \alpha_{i0} = -\frac{T^{-1}\sum_{t=1}^T v_{it}}{E\left[v_{it}^{\alpha_i}\right]} + \frac{\mathcal{B}_i}{T} + o_p\left(\frac{1}{T}\right)$$

where

$$\mathcal{B}_i = \left(E\left[v_{it}^{\alpha_i}\right]\right)^{-2}\left(E\left[v_{it}^{\alpha_i}v_{it}\right] - \frac{E\left[v_{it}^{\alpha_i\alpha_i}\right]E\left[v_{it}^2\right]}{2E\left[v_{it}^{\alpha_i}\right]}\right).$$

and assume that orders in probability correspond to orders in expectation.

- Finally, to obtain B we expand $(nT)^{-1}\sum_{i=1}^n\sum_{t=1}^T u_{it}\left(\widehat{\theta}, \widehat{\alpha}_i\left(\widehat{\theta}\right)\right)$ around θ_0.

- This bias correction formula does not depend on the likelihood setting, and so would be valid for any fixed effects m-estimator with independent observations.

- However, in the likelihood setting because of the information and Bartlett identities we can alternatively write

$$b_i(\theta_0) = \frac{E\left[U_{it}V_{2it}\right]}{2E\left[v_{it}^{\alpha_i}\right]}.$$

- In Example 1 with $\theta = \sigma^2$, we can see that

$$
\begin{aligned}
u_{it} &= -\frac{1}{2\theta_0} + \frac{\left(y_{it} - \alpha_i\right)^2}{2\theta_0^2}, \quad v_{it} = \frac{y_{it} - \alpha_{i0}}{\theta_0}, \quad E\left[v_{it}^{\alpha_i}\right] = -\frac{1}{\theta_0} \\
E\left[u_{it} v_{it}\right] &= 0, \quad U_{it} = u_{it} = -\frac{1}{2\theta_0} + \frac{\left(y_{it} - \alpha_{i0}\right)^2}{2\theta_0^2}, \\
E\left[\mathcal{I}_i\right] &= \frac{1}{2\theta_0^2}, \quad V_{2it} = \frac{\left(y_{it} - \alpha_{i0}\right)^2}{\theta_0^2} - \frac{1}{\theta_0}, \\
E\left[U_{it} V_{2it}\right] &= \frac{1}{\theta_0^2}, \quad \frac{E\left[U_{it} V_{2it}\right]}{E\left[v_{it}^{\alpha_i}\right]} = -\frac{1}{\theta_0}, \\
B &= -\frac{1}{2}\left(\frac{1}{2\theta_0^2}\right)^{-1} \frac{1}{\theta_0} = -\theta_0,
\end{aligned}
$$

  and we obtain

$$
\widehat{\theta}^1 = \widehat{\theta} - \frac{\widehat{B}}{T} = \frac{T+1}{T}\widehat{\theta}.
$$

- Recall that $\widehat{\theta} = \theta_0 - \frac{1}{T}\theta_0 + o_p(1)$ as $n \to \infty$ with T fixed. It follows that

$$
\widehat{\theta}^1 = \theta_0 - \frac{1}{T^2}\theta_0 + o_p(1),
$$

  which shows that the bias of order T^{-1} is removed.

- In fact, $\widehat{\theta}^1 = \widetilde{\theta} - \frac{1}{T^2}\widetilde{\theta}$ where $\widetilde{\theta} = \frac{T}{T-1}\widehat{\theta}$ is the unbiased estimator of θ_0.

Estimators of the Bias

- An estimator of the bias term can be formed using a sample counterpart of the previous formulae. One possibility is

$$\widehat{B}(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{\mathcal{I}}_i \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \widehat{b}_i(\theta)$$

  where

$$\widehat{\mathcal{I}}_i = -\left(\widehat{E}_T \left[\widehat{u}_{it}^{\theta} \right] - \widehat{E}_T \left[\widehat{u}_{it}^{\alpha_i} \right] \widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right]^{-1} \widehat{E}_T \left[\widehat{u}_{it}^{\alpha_i \prime} \right] \right)$$

$$\widehat{b}_i(\theta) = \frac{\widehat{E}_T \left[\widehat{v}_{it} \widehat{v}_{it}^{\alpha_i} \right] \widehat{E}_T \left[\widehat{u}_{it}^{\alpha_i} \right]}{\left(\widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right] \right)^2} - \frac{\widehat{E}_T \left[\widehat{v}_{it} \widehat{u}_{it}^{\alpha_i} \right]}{\widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right]} - \frac{\widehat{E}_T \left[\widehat{v}_{it}^{2} \right]}{2 \widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right]} \left(\frac{\widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i \alpha_i} \right] \widehat{E}_T \left[\widehat{u}_{it}^{\alpha_i} \right]}{\left(\widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right] \right)^2} - \frac{\widehat{E}_T \left[\widehat{u}_{it}^{\alpha_i \alpha_i} \right]}{\widehat{E}_T \left[\widehat{v}_{it}^{\alpha_i} \right]} \right)$$

  where $\widehat{E}_T(.) = \sum_{t=1}^{T} (.) / T$, $\widehat{u}_{it}^{\theta} = u_{it}^{\theta}(\theta, \widehat{\alpha}_i(\theta))$, etc.

- The bias corrected estimator can then be formed with $\widehat{B} = \widehat{B}\left(\widehat{\theta}_T\right)$.

- The other possibility exploits the likelihood setting to replace some derivatives by outer product terms:

$$\widetilde{B}(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{I}}_i\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{b}_i(\theta)$$

where

$$\widetilde{\mathcal{I}}_i = -\widehat{E}_T\left(\widehat{U}_{it}\widehat{U}_{it}'\right), \quad \widehat{U}_{it} = u_{it}(\theta, \widehat{\alpha}_i(\theta)) - \frac{\widehat{E}_T[\widehat{u}_{it}\widehat{v}_{it}]}{\widehat{E}_T[\widehat{v}_{it}^2]}v_{it}(\theta, \widehat{\alpha}_i(\theta)),$$

and

$$\widetilde{b}_i(\theta) = \frac{\sum_{t=1}^{T}\widehat{U}_{it}V_{2it}(\theta, \widehat{\alpha}_i(\theta))}{2\sum_{t=1}^{T}v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))},$$

so that an alternative bias correction can be formed with $\widetilde{B} = \widetilde{B}\left(\widehat{\theta}_T\right)$.

Infinitely Iterated Bias-Correction

- If $\widehat{\theta}$ is heavily biased and it is used in the construction of $\widehat{B}$, it may adversely affect the properties of $\widehat{\theta}^{1}$.

- One way to deal with this problem is to use $\widehat{\theta}^{1}$ in the construction of another $\widehat{B}$, and then form a new bias corrected estimator.

- This procedure could be iterated: Let $\widetilde{B}(\theta)$ denote an estimator of B depending on θ, and suppose that $\widehat{B} = \widetilde{B}\left(\widehat{\theta}\right)$. Then $\widehat{\theta}^{1} = \widehat{\theta} - \widetilde{B}\left(\widehat{\theta}\right)/T$. Iterating gives
$$\widehat{\theta}^{k} = \widehat{\theta} - \widetilde{B}\left(\widehat{\theta}^{k-1}\right)/T, \quad (k = 2, 3, ...).$$

- If this estimator were iterated to convergence, it would give $\widehat{\theta}^{\infty}$ solving
$$\widehat{\theta}^{\infty} = \widehat{\theta} - \widetilde{B}\left(\widehat{\theta}^{\infty}\right)/T.$$

- In general this estimator will not have improved asymptotic properties, but may have lower bias for small T.

- In Example 1 with $\theta_0 = \sigma_0^2$, we can see that
$$\widehat{\theta}^{k} = \frac{T^k + T^{k-1} + \ldots + 1}{T^k}\widehat{\theta} = \frac{T^{k+1} - 1}{T^k (T-1)}\widehat{\theta} \rightarrow \frac{T}{T-1}\widehat{\theta} = \widehat{\theta}^{\infty}$$
as $k \rightarrow \infty$, and the limit $\widehat{\theta}^{\infty}$ has zero bias.

3. Bias-Correction of the Moment Equation

- Another approach to bias correction for fixed effects is to construct the estimator as the solution to a bias corrected version of the first-order conditions.
- Let us consider $\widehat{u}_i(\theta) = \sum_{t=1}^{T} u_{it}(\theta, \widehat{\alpha}_i(\theta))/T$, so that the fixed effects estimator solves $\sum_{i=1}^{n} \widehat{u}_i\left(\widehat{\theta}\right) = 0$. The incidental parameters bias arises because $E\left[\widehat{u}_i(\theta_0)\right] \neq 0$. In fact,

$$E\left[\widehat{u}_i(\theta_0) - \frac{1}{T} b_i(\theta_0)\right] = o\left(\frac{1}{T}\right).$$

- A score-corrected estimator is obtained by solving the modified moment equation

$$\sum_{i=1}^{n}\left[\widehat{u}_i(\theta) - \frac{1}{T}\widehat{b}_i(\theta)\right] = 0$$

where $\widehat{b}_i(\theta)/T$ denotes an estimator of the $1/T$ bias of the expected score, as discussed above. Such an estimator will be expected to be less biased than the MLE $\widehat{\theta}$.

- Alternatively, the bias can be estimated using the estimator of the bias $\widetilde{b}_i(\theta)$ that exploits Bartlett identities, leading to the moment equation

$$\sum_{i=1}^{n}\left[\widehat{u}_i(\theta) - \frac{1}{T}\widetilde{b}_i(\theta)\right] = 0.$$

- The first expression would be valid for any fixed effects m-estimator, whereas the second is appropriate in a likelihood setting (Hahn and Newey, 2004).

- In a likelihood setting it is also possible to form an estimate of $b_i(\theta)$ that uses expected rather than observed quantities, giving rise to alternative score-corrected estimators, such as those considered by Carro (2004) and Fernández-Val (2005) for binary choice.

- To see a connection between bias-correction of the moment equation and iterated bias-correction of the estimator, note that $\widehat{\theta}^{\infty}$ solves the equation

$$\widehat{\theta} - \theta = \frac{1}{T}\widetilde{B}(\theta)$$

or

$$\sum_{i=1}^{n}\left[\widetilde{\mathcal{I}}_i(\theta)\left(\widehat{\theta} - \theta\right) - \frac{1}{T}\widetilde{b}_i(\theta)\right] = 0$$

where $\widetilde{B}(\theta)$ is as in our previous formulae.

- This equation can be regarded as an approximation to a corrected moment equation as long as $\widetilde{\mathcal{I}}_i(\theta)$ is an estimator of $\partial E\left[\widehat{u}_i(\theta)\right]/\partial\theta$ and $\widetilde{b}_i(\theta)/T$ is an estimator of the $1/T$ score bias.

- Thus, the bias-correction of the moment equation can be loosely understood to be an infinitely iterated bias-correction of the estimator.

4. Bias-Correction of the Concentrated Likelihood

- Due to the noise of estimating $\widehat{\alpha}_i(\theta)$, the expectation of the concentrated likelihood is not maximized at the true value of the parameter.
- Let $\ell_i(\theta, \alpha) = \sum_{t=1}^{T} \ell_{it}(\theta, \alpha) / T$ where $\ell_{it}(\theta, \alpha) = \log f(y_{it} \mid \theta, \alpha)$, and let $\overline{\alpha}_i(\theta) = \text{argmax}_\alpha \, \text{plim}_{T \to \infty} \ell_i(\theta, \alpha)$, so that $\overline{\alpha}_i(\theta_0) = \alpha_{i0}$.
- Following Severini (2000) and Pace and Salvan (2004), the concentrated likelihood for unit i
$$\widehat{\ell}_i(\theta) = \ell_i(\theta, \widehat{\alpha}_i(\theta))$$
can be regarded as an estimate of the unfeasible concentrated likelihood:
$$\overline{\ell}_i(\theta) = \ell_i(\theta, \overline{\alpha}_i(\theta)) .$$
- The function $\overline{\ell}_i(\theta)$ is a proper log likelihood which assigns data a density of occurrence according to values of θ and values of the effects along the curve $\overline{\alpha}_i(\theta)$.
- It is a "least-favorable target log likelihood" in the sense that the expected information for θ calculated from $\overline{\ell}_i(\theta)$ coincides with the partial expected information for θ.
- $\overline{\ell}_i(\theta)$ has the usual log likelihood properties: it has zero mean expected score, it satisfies the information matrix identity, and is maximized at θ_0.

- Expanding $\ell_i(\theta, \widehat{\alpha}_i(\theta))$ around $\overline{\alpha}_i(\theta)$ for fixed θ, using a stochastic expansion for $\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)$, and taking expectations, we obtain

$$E\left[\ell_i(\theta, \widehat{\alpha}_i(\theta)) - \ell_i(\theta, \overline{\alpha}_i(\theta))\right] \approx \frac{1}{2} H_i(\theta) Var\left[\widehat{\alpha}_i(\theta)\right] \approx \frac{\beta_i(\theta)}{T}$$

where

$$\beta_i(\theta) = \frac{1}{2} H_i(\theta) Var\left(\sqrt{T}\left[\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)\right]\right) = \frac{1}{2} H_i^{-1}(\theta) \Upsilon_i(\theta)$$

and

$$H_i(\theta) = -E\left[\frac{\partial v_{it}(\theta, \overline{\alpha}_i(\theta))}{\partial \alpha}\right], \quad \Upsilon_i(\theta) = E\left\{\left[v_{it}(\theta, \overline{\alpha}_i(\theta))\right]^2\right\}.$$

- Thus, we expect that

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta)) - \sum_{i=1}^{n} \beta_i(\theta)$$

is a closer approximation to the target log likelihood than $\sum_{i=1}^{n} \sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta))$.

- Letting $\widehat{\beta}_i(\theta)$ be an estimated bias, we then expect an estimator $\widetilde{\theta}$ that solves

$$\widetilde{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \left[\sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta)) - \widehat{\beta}_i(\theta)\right]$$

to be less biased than the MLE $\widehat{\theta}$.

- We can consistently estimate $\beta_i(\theta)$ by

$$\widehat{\beta}_i(\theta) = \frac{1}{2}\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial v_{it}(\theta, \widehat{\alpha}_i(\theta))}{\partial \alpha}\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\left[v_{it}(\theta, \widehat{\alpha}_i(\theta))\right]^2.$$

- Using this form of $\widehat{\beta}_i(\theta)$, $\widetilde{\theta}$ solves the first-order conditions

$$\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} u_{it}(\theta, \widehat{\alpha}_i(\theta)) - \sum_{i=1}^{n}\frac{\partial \widehat{\beta}_i(\theta)}{\partial \theta} = 0.$$

- It can be seen that

$$\frac{\partial \widehat{\beta}_i(\theta)}{\partial \theta} = \widehat{b}_i(\theta).$$

Therefore, the first-order conditions for $\widetilde{\theta}$ and the corresponding bias corrected moment are identical.

- In the likelihood context, we can also consider a local version of the estimated bias constructed as an expansion of $\widehat{\beta}_i(\theta)$ at θ_0 using that at the truth $H_i^{-1}(\theta_0)\Upsilon_i(\theta_0) = 1$:

$$\widehat{\beta}_i(\theta) = \widetilde{\beta}_i(\theta) + O\left(\frac{1}{T}\right)$$

where

$$\widetilde{\beta}_i(\theta) = -\frac{1}{2}\log\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial v_{it}(\theta, \widehat{\alpha}_i(\theta))}{\partial \alpha}\right) + \frac{1}{2}\log\left\{\frac{1}{T}\sum_{t=1}^{T}\left[v_{it}(\theta, \widehat{\alpha}_i(\theta))\right]^2\right\}.$$

- This form of the estimated bias leads to the modified concentrated likelihood

$$\ell_i\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)+\frac{1}{2}\log\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial v_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)}{\partial\alpha}\right]\right\}-\frac{1}{2}\log\left\{\frac{1}{T}\sum_{t=1}^{T}\left[v_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right]^2\right\}.$$

  (cf. DiCiccio and Stern, 1993; Pace and Salvan, 2004).

- This function is maximized at $\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\overline{y}_i\right)^2$ in Example 1.

- It can be easily shown that

$$\frac{\partial\widetilde{\beta}_i\left(\theta\right)}{\partial\theta}=\frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}{\left(-\widehat{E}_T\left[\widehat{v}_{it}^2\right]\right)}\widehat{b}_i\left(\theta\right).$$

  Therefore, the DiCiccio–Stern first-order condition is using a valid estimate of the concentrated score $1/T$ bias as long as the information identity holds.

- In the likelihood setting it is also possible to form estimates of $H_i\left(\theta\right)$ and $\Upsilon_i\left(\theta\right)$ that use expected rather than observed quantities (Severini, 2000).

5. Other Approaches Leading to Bias Correction

- The incidental parameters problem in panel data models can be broadly viewed as a problem of inference in the presence of many nuisance parameters.
- The leading statistical approach under this circumstance has been to search for suitable modification of conditional or marginal likelihoods.
- The modified profile likelihood of Barndorff-Nielsen (1983) and the approximate conditional likelihood of Cox and Reid (1987) belong to this category.
- However, the Barndorff-Nielsen formula is not generally operational, and the one in Cox and Reid requires the availability of an orthogonal effect.

5.1 Approaches Based on Orthogonality

Cox and Reid's Adjusted Profile Likelihood Approach

- Cox–Reid (1987) considered the general problem of inference for a parameter of interest in the presence of nuisance parameters. They proposed an adjustment to the concentrated likelihood to take account of the estimation of the nuisance parameters.

- Their formulation required information orthogonality between the two types of parameters (i.e. that the information matrix be block diagonal).

- In general, the information matrix for (θ, α_i) will not be block-diagonal, although it may be possible to reparameterize α_i as a function of θ and some η_i such that the information matrix for (θ, η_i) is block-diagonal.

- In the panel context, the Cox-Reid (1987) approach maximizes
$$\sum_{i=1}^{n}\sum_{t=1}^{T} \ell_{it}\left(y_{it};\theta,\widehat{\alpha}_i\left(\theta\right)\right) - \frac{1}{2}\sum_{i=1}^{n}\log\left(-\sum_{t=1}^{T}\frac{\partial^2 \ell_{it}\left(y_{it};\theta,\widehat{\alpha}_i\left(\theta\right)\right)}{\partial\alpha_i^2}\right).$$
which was derived as an approximation to the likelihood conditioned on $\widehat{\alpha}_i\left(\theta\right)$.

- This was motivated by the fact that in an exponential family model, it is optimal to condition on sufficient statistics for the nuisance parameters α_i, and these can be regarded as the MLE of α_i chosen to be orthogonal to θ.

- For other problems the idea was to derive a concentrated likelihood for θ conditioned on $\widehat{\alpha}_i\left(\theta\right)$, having ensured via orthogonality that $\widehat{\alpha}_i\left(\theta\right)$ changes slowly with θ.

Relation to Bias-Correction of the Moment Equation

- The first order condition corresponding to the adjusted profile likelihood is equal to

$$\sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\widetilde{b}_i^{CR}\left(\theta\right)\right]=0$$

where

$$\widetilde{b}_i^{CR}\left(\theta\right)=\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{u}^{\alpha_i\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}-\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i\alpha_i}\right]\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\left(\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]\right)^2}.$$

- Ferguson, Reid, and Cox (1991) showed that under orthogonality the expected moment equation has a bias of a smaller order of magnitude than the expected ML score.
- Under information orthogonality $E\left[u_{it}^{\alpha_i}\right]=0$ and $E\left[v_{it}u_{it}^{\alpha_i}\right]=-E\left[u_{it}^{\alpha_i\alpha_i}\right]$. Using these facts and the information identity, the score bias formula becomes

$$b_i\left(\theta_0\right)=\frac{1}{2}\frac{E\left[u_{it}^{\alpha_i\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]}.$$

- Comparison with the Cox–Reid moment adjustment $\widetilde{b}_i^{CR}\left(\theta\right)$ reveals that the latter has an extra term whose population counterpart is equal to zero under orthogonality.

Relation to Bias-Correction of the Concentrated Likelihood

- To see the connection between the Cox–Reid's adjustment, which requires orthogonality, and the one derived from the bias-reduction perspective, which does not, note that the latter can be written as

$$\ell_i\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{2}\log\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial v_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)}{\partial\alpha}\right]\right\}-\frac{1}{2}\log\widehat{\mathrm{Var}}\left(\sqrt{T}\left(\widehat{\alpha}_i\left(\theta\right)-\overline{\alpha}_i\left(\theta\right)\right)\right)$$

  where

$$\widehat{\mathrm{Var}}\left(\sqrt{T}\left(\widehat{\alpha}_i\left(\theta\right)-\overline{\alpha}_i\left(\theta\right)\right)\right)=\frac{T\sum_{t=1}^{T}\left[v_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right]^2}{\left(\sum_{t=1}^{T}\left[v_{it}^{\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right]\right)^2}.$$

- Thus, such a criterion can be regarded as a generalized Cox–Reid adjusted likelihood with an extra term given by an estimate of the variance of $\sqrt{T}\left(\widehat{\alpha}_i\left(\theta\right)-\overline{\alpha}_i\left(\theta\right)\right)$, which accounts for nonorthogonality (Pace and Salvan, 2004).

- Under orthogonality the extra term is irrelevant because the variance of $\widehat{\alpha}_i\left(\theta\right)$ does not change much with θ.

Other Features of the Adjusted Likelihood Approach

- The Cox and Reid's (1987) proposal and other methods in the same literature were not developed to explicitly address the incidental parameter problem in panel data context. Rather, they were concerned with inference in models with many nuisance parameters.
- Neither were developed for the sole purpose of correcting estimation bias, but with the ambitious goal of making the adjusted concentrated likelihood behave like a proper likelihood.
- We can see that the Cox–Reid approach achieves some of these other goals in the context of Example 1, where it can be shown that the unbiased estimator

$$\widehat{\theta} = \frac{1}{n\,(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \overline{y}_i)^2$$

  maximizes the Cox–Reid criterion, and its second derivative delivers $\frac{2\theta^2}{n(T-1)}$ as the estimated variance of $\widehat{\theta}$.
- It is not clear whether such success is specific to the particular example, or not. More complete analysis of other aspects of inference such as variance estimation is beyond the scope of this paper.

Lancaster's (2002) Bayesian Inference

- Lancaster proposed a method of Bayesian inference that is robust to the incidental parameters problem, which like Cox–Reid hinges on the availability of orthogonality.
- In a Bayesian setting, fixed effects are integrated out of the likelihood with respect to a prior distribution conditional on the common parameters (and covariates) $\pi(\alpha \mid \theta)$. In this way, we get an integrated (or random effects) log likelihood of the form
$$\ell_i^I(\theta) = \log \int e^{T\ell_i(\theta,\alpha)} \pi(\alpha \mid \theta)\, d\alpha.$$
- The problem with inferences from $\ell_i^I(\theta)$ is that they depend on the choice of prior for the effects, and are not in general consistent with T fixed.
- In general, the maximizer of $\sum_i \ell_i^I(\theta)$ has a bias of order $O(1/T)$ regardless of $\pi(\alpha \mid \theta)$. However, if α and θ are information orthogonal, the bias can be reduced to $O\left(1/T^2\right)$.
- Lancaster (2002) proposes to integrate out the fixed effects by using a noninformative prior, and use the posterior mode as an estimate of θ.
- The idea is to rely on prior independence between fixed effects and θ, having chosen an orthogonal reparameterization, say $\alpha_i = \alpha(\theta, \eta_i)$, that separates the common parameter θ from the fixed effects η_i in the information matrix sense.

- His estimator $\widehat{\theta}_L$ takes the form

$$\widehat{\theta}_L = \underset{\theta}{\operatorname{argmax}} \int \cdots \int \prod_{i=1}^{n} \prod_{t=1}^{T} f\left(y_{it} \middle| \theta, \alpha\left(\theta, \eta_i\right)\right) d\eta_1 \cdots d\eta_n.$$

- In Example 1 with $\theta = \sigma^2$, we have $E\left[u_{it} v_{it}\right] = 0$ so the reparameterization is unnecessary. Lancaster's estimator would therefore maximize

$$\begin{aligned} &\int \cdots \int \prod_{i=1}^{n} \prod_{t=1}^{T} \frac{1}{\sqrt{\theta}} \exp\left(-\frac{\left(y_{it} - \alpha_i\right)^2}{2\theta}\right) d\alpha_1 \cdots d\alpha_n \\ &\propto \frac{1}{\left(\sqrt{\theta}\right)^{T-1}} \exp\left(-\frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \overline{y}_i\right)^2}{2\theta}\right), \end{aligned}$$

leading to

$$\widehat{\theta}_L = \frac{1}{n\left(T-1\right)} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \overline{y}_i\right)^2,$$

which has no bias.

- The asymptotic properties of $\widehat{\theta}_L$ are not yet fully worked out. It is in general expected $\widehat{\theta}_L$ removes bias only up to $O\left(T^{-1}\right)$, although we can find examples where $\widehat{\theta}_L$ eliminates bias of even higher order.

5.2 Overcoming Infeasibility of Orthogonalization

- The Cox–Reid and Lancaster approaches are successful only when the parameter of interest can be orthogonalized with respect to the nuisance parameters.
- In general, such reparameterization requires solving some partial differential equations, and the solution may not exist. Because parameter orthogonalization is not feasible in general, such approach cannot be implemented for arbitrary models.
- This problem can be overcome by adjusting the moment equation instead of the concentrated likelihood. We discuss two approaches in this regard, one introduced in Woutersen (2002) and the other in Arellano (2003).
- We will note that these two approaches result in identical estimators.

Woutersen's (2002) Approximation

- Woutersen (2002) provided an insight on the role of Lancaster's posterior calculation in reducing the bias of the fixed effects.
- Assume for simplicity that the common parameter θ is orthogonal to α_i in the information sense, and no reparameterization is necessary to implement Lancaster's proposal.
- Given the posterior
$$\prod_{i=1}^{n}\left(\int\prod_{t=1}^{T}f\left(y_{it}\right|\theta,\alpha_i)\,d\alpha_i\right),$$
the first order condition that characterize the posterior mode can be written as
$$0=\sum_{i=1}^{n}\frac{\int\left(\sum_{t=1}^{T}u_{it}\left(\theta,\alpha_i\right)\right)\prod_{t=1}^{T}f\left(y_{it}\right|\theta,\alpha_i)\,d\alpha_i}{\int\prod_{t=1}^{T}f\left(y_{it}\right|\theta,\alpha_i)\,d\alpha_i}.$$
- Woutersen pointed out that the ith summand on the right can be approximated by
$$\sum_{t=1}^{T}\widehat{u}_{it}-\frac{1}{2}\frac{\sum_{t=1}^{T}\widehat{u}_{it}^{\alpha_i\alpha_i}}{\sum_{t=1}^{T}\widehat{v}_{it}^{\alpha_i}}+\frac{1}{2}\frac{\left(\sum_{t=1}^{T}\widehat{v}_{it}^{\alpha_i\alpha_i}\right)\left(\sum_{t=1}^{T}\widehat{u}_{it}^{\alpha_i}\right)}{\left(\sum_{t=1}^{T}\widehat{v}_{it}^{\alpha_i}\right)^2}.$$

- Therefore, his estimator under parameter orthogonality is the solution to
$$0=\sum_{i=1}^{n}\left[\sum_{t=1}^{T}u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{u}^{\alpha_i\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}+\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i\alpha_i}\right]\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\left(\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]\right)^2}\right], \tag{1}$$
which coincides with the Cox–Reid moment equation.

- Woutersen pointed out that the moment function
$$\overline{u}_{it}\left(\theta,\alpha\right)\equiv u_{it}\left(\theta,\alpha\right)-\rho_i\left(\theta,\alpha\right)v_{it}\left(\theta,\alpha\right)$$
where
$$\rho_i\left(\theta,\alpha\right)\equiv\frac{\int u_i^{\alpha}\left(y;\theta,\alpha\right)f_i\left(y;\theta,\alpha\right)dy}{\int v_i^{\alpha}\left(y;\theta,\alpha\right)f_i\left(y;\theta,\alpha\right)dy}$$
would satisfy the orthogonality requirement in the sense that at true values
$$E\left[\overline{u}_{it}^{\alpha}\left(\theta_0,\alpha_{i0}\right)\right]=0.$$

- Note that $\overline{u}_{it}\left(\theta,\alpha_i\right)$ can be regarded as a feasible version of $U_{it}\left(\theta,\alpha_i\right)$.

- Woutersen's moment equation when parameter orthogonality is unavailable is therefore obtained by replacing $u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)$ in (1) by $\overline{u}_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)$.

Arellano's (2003) Proposal

- Arellano (2003) assumes that parameter orthogonalization is feasible under certain reparameterization, and that the Cox–Reid moment equation is written for such a reparameterized model.

- He then proposes to rewrite the moment equation in terms of the original parameterization, and obtains
$$0 = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} u_{it} \left(\theta, \widehat{\alpha}_i \left(\theta \right) \right) - \widetilde{b}_i^{CR} \left(\theta \right) + \frac{\partial \rho_i \left(\theta, \alpha \right)}{\partial \alpha} \, |_{\alpha = \widehat{\alpha}_i(\theta)} \right],$$
after suppressing a transformation specific term that is irrelevant for bias reduction.

- This moment equation turns out to be identical to Woutersen's (2002) equation.

- Arellano's derivation was based on the implicit assumption that parameter orthogonalization is feasible. Our discussion suggests that this is not necessary. After all, his procedure is identical to Woutersen's, which does not require orthogonality.

- Indeed, Carro (2004) has shown that such moment equation reduces the order of the score bias regardless of the existence of an information orthogonal reparameterization.

Relation to Bias-Correction of the Moment Equation

- The moment equation used by Woutersen, Arellano, and Carro can be written as

$$\sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\widetilde{b}_i^{W}\left(\theta\right)\right]=0$$

where

$$\widetilde{b}_i^{W}\left(\theta\right)=\widetilde{b}_i^{CR}\left(\theta\right)-\frac{\partial\rho_i\left(\theta,\alpha\right)}{\partial\alpha}\Big|_{\alpha=\widehat{\alpha}_i(\theta)}.$$

- Comparing the resulting expression with the theoretical bias, we note that this moment condition is using a valid estimate of the $1/T$ score bias as long as the information identity holds, so that in general it will be appropriate in likelihood settings.

- The estimated bias $\widetilde{b}_i^{W}\left(\theta\right)$ uses a combination of observed and expected terms, and contrary to the situation under orthogonality, there is no redundant term.

- The term $\partial\rho_i\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)/\partial\alpha$ can be interpreted as a measure of how much the variance of $\widehat{\alpha}_i\left(\theta\right)$ changes with θ.

- In this respect, we note the equivalence between the derivative of the log variance of $\widehat{\alpha}_i\left(\theta\right)$ and a sample counterpart of $\partial\rho_i\left(\theta_0,\alpha_{i0}\right)/\partial\alpha$.

6. QMLE for Dynamic Models

- The starting point of our discussion so far has been the assumption that the fixed effects estimator actually maximizes the likelihood. When we defined $\widehat{\theta}$ to be a maximizer of

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \log f\left(y_{it} \mid x_{it}, \theta, \widehat{\alpha}_i\left(\theta\right)\right),$$

  we assumed that

  – (i) xs are strictly exogenous,

  – (ii) ys are independent over t given xs, and

  – (iii) f is the correct (conditional) density of y given x.

- We noted that some of the bias-correction methods did not depend on the likelihood setting, while others, that relied on the information or Bartlett identities, did. However, in all cases assumptions (i) and (ii) were maintained.

- For example, if the binary response model

$$y_{it} = 1\left(x_{it}'\theta + \alpha_i + e_{it} > 0\right)$$

  where the marginal distribution of e_{it} is $\mathcal{N}\left(0, 1\right)$, is such that e_{it} is independent over t, and if it is estimated by nonlinear least squares, our first bias formula is valid.

- In the likelihood setting, assumption (ii) can be relaxed choosing estimates of bias corrections that use expected rather than observed quantities.
  - This is possible because the likelihood fully specifies the dynamics.
  - It is simple if the required expected quantities have closed form expressions, as in the dynamic probit models in Carro (2004) and Fernández-Val (2005).
- In a nonlikelihood setting, our analysis can be generalized to the case when the fixed effects estimator maximizes
$$\sum_{i=1}^{n} \sum_{t=1}^{T} \psi \left(z_{it}; \theta, \widehat{\alpha}_i \left(\theta \right) \right)$$
for an arbitrary ψ under some regularity conditions, thereby relaxing assumptions (i) and (ii).
- For example, the binary response model can still be analyzed by considering the fixed effects probit MLE even when e_{it} has an arbitrary unknown serial correlation.
- The analysis for this more general model gets to be more complicated because estimates of the expectations of products of average derivatives should incorporate arbitrary forms of serial correlation, which was a non-issue in the simpler context.
- Hahn and Kuersteiner (2004) provide an analysis that incorporate such complication.

7. Estimation of Marginal Effects

- It is sometimes of interest to estimate quantities such as
$$\mu = \operatorname{plim} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} m\left(z_{it}; \theta, \alpha_i\right).$$
- For example, it may be of interest to estimate the mean marginal effects
$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \phi\left(x_{it}'\theta + \alpha_i\right) \theta$$
for the binary response model, where ϕ denotes the $\mathcal{N}(0,1)$ density.
- It would be sensible to estimate such quantities by
$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} m\left(z_{it}; \widetilde{\theta}, \widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)$$
where $\widetilde{\theta}$ denotes a bias-corrected version of $\widehat{\theta}$.
- In order to obtain a bias-corrected estimator of μ, it is useful to think about this type of quantity as a solution to the (infeasible) moment equation
$$\sum_{i=1}^{n} \sum_{t=1}^{T} \left(m\left(z_{it}; \widehat{\alpha}_i\left(\theta_0\right)\right) - \widehat{\mu}\right) = 0, \quad \sum_{t=1}^{T} v\left(z_{it}; \widehat{\alpha}_i\left(\theta_0\right)\right) = 0$$
where for simplicity we suppressed the dependence of m on θ.

- Let

$$M\left(z_{it};\alpha_i\right)=m\left(z_{it};\alpha_i\right)-v\left(z_{it};\alpha_i\right)\frac{E\left[m^{\alpha_i}\left(z_{it};\alpha_i\right)\right]}{E\left[v^{\alpha_i}\left(z_{it};\alpha_i\right)\right]}$$

  and note that $\widehat{\mu}$ solves

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left[M\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\mu}\right]=0.$$

- We can bias-correct this moment equation using the same intuition as before.
- We then obtain a bias corrected version of the moment equation

$$\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\left[M\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\widehat{\mu}}\right]+\frac{\widehat{E}_T\left(v_{it}M_{it}^{\alpha_i}\right)}{\widehat{E}_T\left(v_{it}^{\alpha_i}\right)}-\frac{\widehat{E}_T\left(v_{it}^2\right)\widehat{E}_T\left(M_{it}^{\alpha_i\alpha_i}\right)}{2\left[\widehat{E}_T\left(v_{it}^{\alpha_i}\right)\right]^2}\right)=0.$$

- Replacing $M\left(z_{it};\theta_0,\widehat{\alpha}_i\left(\theta_0\right)\right)$ in this moment equation by the feasible version:

$$m\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)-v\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)\frac{\widehat{E}_T\left[m^{\alpha_i}\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)\right]}{\widehat{E}_T\left[v^{\alpha_i}\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)\right]},$$

  we obtain a bias corrected estimator, similar to those discussed in Hahn and Newey (2004), and Fernández-Val (2005).

8. Automatic methods

- We have so far discussed analytic methods of bias correction, but we may be able to bypass such analysis, and rely on numerical methods. We discuss two such procedures.

Panel Jackknife

- Let $\widehat{\theta}_{(t)}$ be the fixed effects estimator based on the subsample excluding the observations of the tth period. The jackknife estimator is

$$\widetilde{\theta} \equiv T\widehat{\theta} - (T-1)\sum\nolimits_{t=1}^{T} \widehat{\theta}_{(t)}/T$$

or

$$\widetilde{\theta} \equiv \widehat{\theta} - \frac{\widetilde{B}}{T}, \qquad \frac{\widetilde{B}}{T} = (T-1)\left(\frac{1}{T}\sum\nolimits_{t=1}^{T} \widehat{\theta}_{(t)} - \widehat{\theta}\right).$$

- To explain the bias correction from this estimator, consider a further expansion

$$\theta_T = \theta_0 + \frac{B}{T} + \frac{D}{T^2} + O\left(\frac{1}{T^3}\right).$$

- The limit of $\widetilde{\theta}$ for fixed T and its change with T shows the effect of the correction:

$$T\theta_T - (T-1)\,\theta_{T-1} = \theta_0 + \left(\frac{1}{T} - \frac{1}{T-1}\right) D + O\left(\frac{1}{T^2}\right) = \theta_0 + O\left(\frac{1}{T^2}\right)$$

or

$$(T-1)\left(\theta_{T-1} - \theta_T\right) = \frac{B}{T} + O\left(\frac{1}{T^2}\right).$$

- Thus, we see that the asymptotic bias of the jackknife estimator is of order $1/T^2$.

- Hahn and Newey (2004) established that $\sqrt{nT}\left(\widetilde{\theta}-\theta_0\right)$ has the same asymptotic variance as $\sqrt{nT}\left(\widehat{\theta}-\theta_0\right)$ when $n/T \rightarrow \rho$. This implies that the bias reduction is achieved without any increase in the asymptotic variance.
- In Example 1, it is straightforward to show that
$$\widetilde{\theta}=\frac{1}{n\left(T-1\right)}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\overline{y}_i\right)^2,$$
so that the jackknife bias correction completely removed bias in this example.
- Another possibility is to use $\widehat{\theta}_{(T)}$ as the sample analog of θ_{T-1}, where $\widehat{\theta}_{(T)}$ is the MLE based on the first $T-1$ observations. It turns out that such procedure will be accompanied by some large increase in variance.
- The panel jackknife is easiest to understand when y_{it} is independent over time. When it is serially correlated, it is not yet clear how it should be modified.

Bootstrap Adjusted Concentrated Likelihood

- Simulation methods can also be used for bias correction of moment equations and objective functions. Pace and Salvan (2004) have suggested a bootstrap approach to adjust the concentrated likelihood.

- Consider generating parametric bootstrap samples $\{y_{i1}(r), ..., y_{iT}(r)\}_{i=1}^{n}$ $(r = 1, ..., R)$ from the models $\left\{\prod_{t=1}^{T} f\left(y_t \mid \widehat{\theta}, \widehat{\alpha}_i\right)\right\}_{i=1}^{n}$ to obtain $\widehat{\alpha}_i^{[r]}(\theta)$ as the solution to

$$\widehat{\alpha}_i^{[r]}(\theta) = \underset{\alpha}{\operatorname{argmax}} \sum\nolimits_{t=1}^{T} \log f\left(y_{it}(r) \mid \theta, \alpha\right) \quad (r = 1, ..., R).$$

- A simulation adjusted log-likelihood for the i-th unit is

$$\bar{\ell}_i^S(\theta) = \frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T} \ell_{it}\left(\theta, \widehat{\alpha}_i^{[r]}(\theta)\right).$$

- Alternatively, Pace and Salvan consider the generalized Cox–Reid form, using a bootstrap estimate of $V_i\left[\widehat{\alpha}_i(\theta)\right]$ given by

$$\widetilde{V}_i\left[\widehat{\alpha}_i(\theta)\right] = \frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\alpha}_i^{[r]}(\theta) - \widehat{\alpha}_i(\theta)\right]^2,$$

  which leads to

$$\bar{\ell}_i^{SA}(\theta) = \sum_{t=1}^{T} \ell_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) - \frac{1}{2}\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial v_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right)}{\partial \alpha}\right)\widetilde{V}_i\left[\widehat{\alpha}_i(\theta)\right].$$

Numerical Illustrations

- Let us consider the probit model $y_{it} = 1\left(\theta_0 x_{it} + \alpha_{i0} + v_{it} \geq 0\right)$ in which $T = 2$ and x_{it} is a time dummy such that $x_{i1} = 0$ and $x_{i2} = 1$.
- Figure 1, taken from Arellano (2003), shows the probability limits of ML and MML for $\mathcal{N}\left(0, \sigma_\alpha^2\right)$ individual effects with $\sigma_\alpha^2 = 0.1, 1,$ and 10, as well as for Cauchy effects.
- Thus we are assessing the value of a large-T adjustment when $T = 2$.
- The impact of changing the distribution of the effects is small for both ML and MML. The adjustment for probit produces a good improvement given that T is only two.

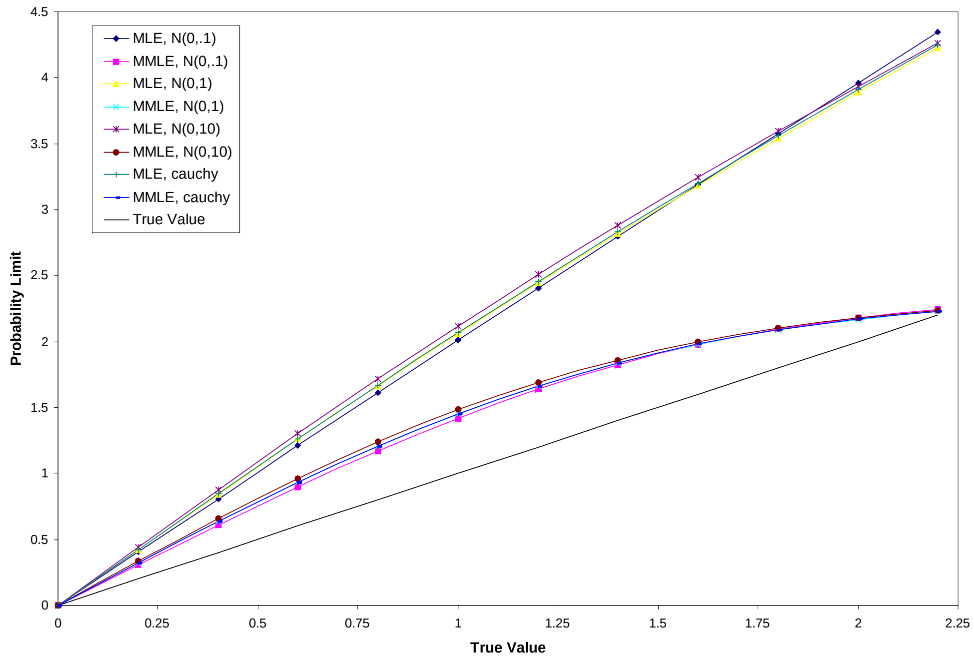

Figure 1: Probability limits for a probit model with $T = 2$

Table 1
Simulations for a dynamic Logit model
taken from J. Carro (2004)

		$T=4$	$T=8$	$T=16$
		$\theta_0 = 0.5$		
MLE	Bias	-2.55	-0.76	-0.31
	MAE	2.55	0.76	0.31
MMLE	Bias	-0.55	-0.11	-0.02
	MAE	0.55	0.13	0.07
HK	Bias	-0.004	-0.05	-0.05
	MAE	0.40	0.13	0.07
		$\theta_0 = 2$		
MLE	Bias		-0.65	-0.30
	MAE		0.65	0.30
MMLE	Bias		-0.23	-0.04
	MAE		0.23	0.08
HK	Bias		-0.20	-0.20
	MAE		0.23	0.20

$y_{it} = 1\left(\theta_0 y_{it-1} + x_{it} + \alpha_{i0} + v_{it} \geq 0\right)$, $\alpha_{i0} = \overline{x}_i$, $x_{it} \sim \mathcal{N}\left(0, \pi^2/3\right)$, v_{it} ∼logistic; 1000 replications; $n = 250$; HK is the Honoré–Kyriazidou estimator.

Concluding Remarks

- We discussed a variety of methods of estimation of nonlinear fixed effects panel data models with reduced bias properties.
  - Alternative approaches to bias correction based on adjusting the estimator, the moment equation, and the criterion function have been considered.
  - We have also discussed approaches relying on orthogonalization and automatic methods, as well as the connections among the various approaches.
- Next in the agenda, it is important to find out how well each of these methods work for specific models and data sets of interest in applied econometrics.
  - In this regard, the Monte Carlo results and empirical estimates obtained by Carro (2004) and Fernández-Val (2005) for binary choice models are very encouraging.
- We have focused on bias reduction, but other theoretical properties should play a role in narrowing the choice of bias-reducing estimation methods.
  - In the likelihood context it is natural to seek an adjusted concentrated likelihood that behaves like a proper likelihood.
  - In this respect, information bias reduction and invariance to reparameterization are relevant properties in establishing the relative merits of different bias-reducing estimators.